Expert Systems with Applications 40 (2013) 7400–7406

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Linear Bayes policy for learning in contextual-bandits

José Antonio Martín H. [a,*], Ana M. Vargas [b]

[a] Computer Architecture and Automation, Universidad Complutense de Madrid, Spain
[b] Industrial Engineering, Business Administration and Statistics, Universidad Politécnica de Madrid, Spain

ARTICLE INFO

Keywords:
Contextual bandits
Online advertising
Recommender systems
One-to-one Marketing
Empirical Bayes

ABSTRACT

Machine and Statistical Learning techniques are used in almost all online advertisement systems. The problem of discovering which content is more demanded (*e.g.* receive more clicks) can be modeled as a multi-armed bandit problem. Contextual bandits (*i.e.*, bandits with covariates, side information or associative reinforcement learning) associate, to each specific content, several features that define the "context" in which it appears (*e.g.* user, web page, time, region). This problem can be studied in the stochastic/statistical setting by means of the conditional probability paradigm using the Bayes' theorem. However, for very large contextual information and/or real-time constraints, the exact calculation of the Bayes' rule is computationally infeasible. In this article, we present a method that is able to handle large contextual information for learning in contextual-bandits problems. This method was tested in the Challenge on Yahoo! dataset at ICML2012's Workshop "new Challenges for Exploration & Exploitation 3", obtaining the second place. Its basic exploration policy is deterministic in the sense that for the same input data (as a time-series) the same results are obtained. We address the deterministic exploration vs. exploitation issue, explaining the way in which the proposed method deterministically finds an effective dynamic trade-off based solely in the input-data, in contrast to other methods that use a random number generator.

1. Introduction

In statistical decision theory, conditional probability through the Bayes' theorem provides an optimal decision rule. However, in sequential decision problems under stochastic conditions, when informed decisions must be computed sequentially with larger previously unknown side information, or the time to take decisions is critical, the exact calculation of the Bayes' optimal rule is computationally intractable, or –even worse– non-computable. Nowadays, this is an increasingly common picture and so empirical approximations to Bayesian methods (Robbins, 1964) are finding great applicability in the Machine and Statistical Learning fields.

One of the most general and theoretically sound approaches in Machine Learning is the Inductive Inference Theory, where the learner tries to *predict* future events from the history (sequence) of past events (see Solomonoff, 1964; Vovk, Gammerman, & Shafer, 2005), that is, "learning from observations". However, in general, there is no way to predict a sequence completely. Thus one can divide a sequence in two parts: all that can be predicted, and all that cannot. Moreover, even for the predictable part some kind of search or trial–error process is often required. This effort is what is called *exploration* and deciding when one should explore or just use the current knowledge to make an educated guess (i.e., *exploit* current knowledge) is the so called *exploration vs. exploitation problem* (see Holland, 1975; March, 1991 or Kaelbling, Littman, & Moore, 1996). The balance, ratio or proportion between these two opposed alternatives is called *the exploration/ exploitation trade-off* and is performed following an arbitrary (ad hoc) algorithm referred to as the *exploration policy*.

In this article, we present a (empirical) Bayes-like method for learning in contextual-bandits problems. This method is able to take effective advantage of a large contextual information in an efficient manner. In addition, one key-feature is the autonomous exploration / exploitation trade-off that it achieves deterministically, based solely in the input-data, in contrast to other methods that use a random number generator.

1.1. Bandit problems and the exploration vs. exploitation problem

The *multi-armed bandit problem* (MAB) is the sequential decision task where an agent (gambler or player) must decide or choose (pull) a set of actions (arms) to take at each time step, by following some informed strategy (a policy). For each chosen action, the agent receives a corresponding numerical *payoff* (reward) following an unknown probability distribution that may evolve in

* Corresponding author. Address: Facultad de Informática, Universidad Complutense de Madrid, C. Prof. José García Santesmases, s/n, 28040 Madrid, Spain. Tel.: +34 91 394 764; fax: +34 91 394 7510.
E-mail addresses: jamartinh@fdi.ucm.es (J.A. Martín H.), ana.vargas@upm.es (A.M. Vargas).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.07.041

time for each different arm. Then, the agent can use such payoffs to improve its selection strategy to decide the choice of actions in the future to maximize the cumulative payoff in the long-run. Therefore, this becomes the problem of estimating the payoff-probability of each arm (over time).

A standard way of analyzing this maximization problem is to define it in terms of the minimization of the loss or *regret* with respect to the *optimal-policy* that always plays the best arm $a^*(t)$ at trial t (Auer, Cesa-Bianchi, & Fischer, 2002). Hence, a natural measure of optimality in terms of the regret $R_A(T)$ can be expressed in the following way:

$$R_A(T) \stackrel{\text{def}}{=} \mathbf{E}\Big[\sum_{t=1}^{T} r(t,a^*)\Big] - \mathbf{E}\Big[\sum_{t=1}^{T} r(t,a)\Big], \quad (1)$$

where A is the current algorithm and $r(t,a)$ the payoff obtained by playing arm a at trial t.

The multi-armed bandit problem was observed and studied by Robbins (1952), as the problem of sequential design of experiments (Wald, 1947), and extensively studied in statistics by Berry and Fristedt (1986). It was formalized and solved optimally by Gittins, Weber, and Glazebrook (1989) for the special case in which the payoffs of all the arms are independent and that only one arm may evolve at each play.

The exploration vs. exploitation dilemma is what makes the MAB specially useful in several disciplines (Berry & Fristedt, 1986; Jun, 2004; Audibert, Munos, & Szepesvári, 2009; Bubeck & Cesa-Bianchi, 2012). That is, finding the right proportion between these two opposed "intentions":

1. Exploit the current knowledge to guess the best choice.
2. Explore an unknown or suboptimal choice to improve the knowledge about the problem (when possible).

Intuitively, your first impulse is to minimize exploration in order to increase the chance of getting a higher payoff, or –conversely– experience a low regret. However, which is the minimum exploration rate to minimize the regret in the long-run? Table 1 show us a subtle clue! The last row gives the relation between the exploration vs. exploitation problem with the data-compression one, for which we know that it is non-computable, *i.e.*, no general lossless compression method may exist.

This follows directly from the non-computability of Kolmogorov's complexity $K(s)$ of a string s. In the general case, we can't encode any sequence S_φ of length $\ell(S_\varphi)$ in a shorter sequence S_ϵ of length $\ell(S_\epsilon) < \ell(S_\varphi)$.

Now, since any *optimal* run of a sequential decision problem φ defines (obviously) a sequence of decisions S^*_φ of length $\ell(S_\varphi)^*$, then we cannot, in general, find a shorter sequence S_ϵ of length $\ell(S_\epsilon) < \ell\left(S^*_\varphi\right)$ that would predict S^*_φ (for any non-trivial S_φ).

Therefore, in general, a shorter sequence that specifies an optimal exploration/ exploitation trade-off that serves to predict the optimal sequence of actions does not exists. Otherwise, we would be able to create a universal lossless compression program by encoding particular playing sequences as strings.

This tell us that there are tasks in which the optimal solution is a pure exploration approach since there will be problems in which learning (and so prediction) is impossible at all. However, despite this bad news, in a sense, this is a full employment theorem for bandits, and so it is possible to find suboptimal exploration policies that significantly improve learning.

1.2. Contextual bandits and online advertising/recommender systems

Nowadays, Machine Learning and statistical techniques are used in almost all online advertisement and recommendation systems (Konstan & Ried, 2012; Agarwal, Chen, Elango, & Ramakrishnan, 2013). The problem of discovering which content is more demanded (*e.g.* receive more clicks), or which product is more likely to be consumed if displayed in an online advertisement system, can be modeled mathematically as a multi-armed bandit problem.

In online advertising, the *click-through rate* (CTR) is an index used to measure purchase propensity. This index is calculated as the proportion obtained by dividing the number of clicks received by an advertising-banner by the number of its impressions or displays (Agarwal, Chen, & Elango, 2009; Wang, Li, Cui, Zhang, & Mao, 2011). From here, a common approach is to model online advertising as a multi-armed bandit problem for maximizing the CTR of the repeated interaction cycle whereby the system selects an article (arm) from a pool, recommends it by displaying the article to a particular user (pull the arm) and then observes whether the user clicked or not the recommended article (get the payoff or reward).

The Contextual bandits model, also known as bandits with covariates, side information, associative bandits or associative reinforcement learning (Langford & Zhang, 2007; Li, Chu, Langford, & Schapire, 2010), or simply the reinforcement learning case when there are multiple states but reinforcement is immediate (Kaelbling et al., 1996), is a natural extension of the multi-armed bandit problem. Contextual bandits incorporates additional information (context) to the decision making process. The assumption is that the payoff obtained by playing an arm, is –up to some degree if not totally– dependent on such contextual information (i.e., a covariate). This kind of problem appears to have wider applicability in practice, since problems that can be solved optimally without considering contextual information are not so common (Langford & Zhang, 2007). For example, feature-based recommender systems in general (Weng & Liu, 2004), and particularly news recommendation systems (Li et al., 2010; Liu, Dolan, & Pedersen, 2010) can be modeled naturally as contextual bandits.

Following the terminology of Li et al. (2010) (with some minor variations): a contextual-bandit algorithm A proceeds in discrete trials $t = 1,2,3,\ldots T$. At each trial t:

1. The algorithm observes a set $A(t)$ of arms (e.g. articles, options, choices) and a features vector $\mathbf{x}(t)$ (the *context*).

Table 1
Some examples of equivalent terms to exploration and exploitation that are used in different fields.

Area or Discipline	Exploration	vs.	Exploitation
Sequential decision making	exploration	vs.	exploitation
Compressed sensing	sensor-reading	vs.	signal-reconstruction
Statistics and Machine Learning	memorizing data	vs.	generalizing
Curve-fitting	acquire-points	vs.	interpolation
Economics	risk-taking	vs.	risk-avoiding
Finance	investing	vs.	saving
Marketing	diversification/proliferation	vs.	concentration strategy
Medicine	experimental treatments	vs.	safety and efficacy
Data-compression	store-data	vs.	space-savings